Birney, D.P., & Beckmann, J.F., Beckmann, N., & Double, K.S. (in press). Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions. *Intelligence*. (10.1016/j.intell.2017.01.005)

Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions

Damian P. Birney[*1]

Jens F. Beckmann[2]

Nadin Beckmann[2]

Kit S. Double[1]

Running Head: Beyond the intellect

Key words: Fluid Intelligence; Learning; Complexity; Confidence Ratings; Self-Regulatory Processes

1 School of Psychology, University of Sydney, Australia

2 School of Education, Durham University, UK

*Corresponding Author
Damian P. Birney
School of Psychology
University of Sydney
Sydney, NSW, 2006, Australia
E-mail: damian.birney@sydney.edu.au

Abstract

The Raven's Progressive Matrices (RPM) test entails a 40-min contextualized interaction with a set of progressively difficult cognitive activities. Item-to-item experiences accumulate to total scores determined by, and reflective of, cognitive abilities. The current research is interested in what happens during those 40 minutes. Personality (Openness, Extraversion and Neuroticism) and metacognitive factors have consistently been associated, albeit at low levels, with performance. 252 industry managers completed, inter alia, the RPM either with or without confidence ratings. Using multi-level modeling and controlling for general ability, we investigate whether a) experiential factors emerge in individual performance trajectories, b) whether trajectories are associated with cognitive and personality factors, and c) whether requirements to externalize metacognitive reflection (provide confidence ratings) links to performance. Results suggest that metacognitive reflection impeded performance; that learning trajectories are separable from performance trajectories; and that trajectories are statistically moderated, most notably by Neuroticism, over and above cognitive ability. Modeling item-level responses following experimental manipulations that serve as a catalyst for modifying cognition-personality relations, provides an important avenue for integrating experimental and differential methods. Psychometric complexity (ψ_C) and psychometric learning (ψ_L) are proposed as theoretically derived empirical bases to ground investigations of statistical moderation. Together they may provide a bridge to causal accounts of the divide between intelligence and personality.

Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions

The Raven's Progressive Matrices (RPM) is a widely known group-administered test of general fluid intelligence (*Gf*). In a standard administration of Set II of the advanced RPM, 36 progressively more complex items are presented within a time limit of 40 minutes. As the test progresses, each successive item demands induction of different rules, multiple rules, and/or more complex instantiations of rules from earlier items (Carpenter, Just, & Shell, 1990). This test of increasingly complex items was Raven's (1941) operationalization of intelligence as the capacity to perceive relations and educe correlates (Spearman, 1927). From a test-taker's perspective, the RPM is a more or less idiosyncratic experience with distinct, challenging cognitive activities lasting over a period of up to 40 minutes. Adopting a multi-level approach, the current research is interested in modeling the moderation of a broad range of between-person differences on within-person performance trajectories spanning that 40 minutes.

As the basis of this, we note that modern conceptualizations of intelligence, cognitive ability, and reasoning as they are realized in everyday experiences, warrant greater attention toward non-cognitive influences (Ackerman, 1988; Sitzmann & Ely, 2011). A range of personality and metacognitive factors has consistently been observed to be associated with intellectual performance (Bandura, 1991; Soubelet & Salthouse, 2011), although the strengths of those associations tend to be rather small for all but a few of these factors. The goal of the current work is to map the influence of personality and metacognitive reflection, not simply on total RPM scores, but also on item-to-item performance trajectories. For reasons to be presented, we focus on Openness, Extraversion, and Neuroticism as moderating personality facets, and metacognitive reflection as operationalized by the requirement to provide item-specific confidence ratings.

The research presented here aims to make a number of important contributions. First, it extends investigations of the cognition-personality links to include item-by-item RPM performance trajectories. Second, it investigates whether being required to externalize metacognitive reflection has an impact on RPM performance. Third, building on Schweizer and colleagues fixed-link SEM models (e.g., Ren, Goldhammer, Moosbrugger, & Schweizer, 2012; Schweizer, 2006), which distinguish item-difficulty from item-position effects, it introduces a theory-driven conceptualization of statistical moderation of performance trajectories. Finally, the research benefits from sampling experienced business managers and exposing them to ecologically valid assessment conditions that matter to the individual. We begin by outlining the rationale for the type of statistical moderation we investigate.

Conceptualizing moderators of RPM performance: Psychometric Complexity

Conceptually, the capacity to learn and to deal with novelty (Crawford, 1991; Sternberg, 1985; Sternberg & Gastel, 1989) and complexity (Marshalek, Lohman, & Snow, 1983; Stankov, 2000) have been regarded as reflecting elementary reasoning abilities central to *Gf* (Carpenter et al., 1990; Primi, 2001). At a task level, increasing novelty or complexity should therefore result in greater demand being placed on *Gf* resources, and hence concomitantly, be associated with a monotonic increase in the correlation between task performance and measures of *Gf* (Stankov & Crawford, 1993). Birney and Bowman (2009) differentiated process-oriented *psychometric complexity* factors from other factors that make solutions difficult but do not necessarily place higher demands on *Gf* (and thus do not result in changes of *Gf*-performance correlations). They investigated *Gf* processes by experimentally manipulating relational processing demands of reasoning tasks (Birney, Halford, & Andrews, 2006; Halford, Wilson, & Phillips, 1998), with the expectation that this would result in a *psychometric complexity effect* – that is, an increase in the correlation between task performance and independent measures of *Gf* as complexity of the items increased. Although relational reasoning overall was correlated with *Gf*, Birney and Bowman found *Gf* better differentiated test takers' capacity to maintain information across multiple steps within an item (i.e., serial processing demand), rather than relational complexity differences between-items. That the psychometric complexity effect was present only as a function of within-problem WM demand is broadly consistent with Engle and colleagues who argue that the pervasive correlation between WM and *Gf* is driven by the capacity for *controlled attention* (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005).

In asking the question of whether the observed progression of RPM item difficulty is a function of a psychometric complexity effect or some other difficulty factor, we need to consider the processes involved. Prima facie, two cognitive factors are most prominent, reasoning and learning. The RPM requires an individual to determine the answer to a particular problem by inductive reasoning with a relatively small set of rules (Carpenter et al., 1990). Although some authors have suggested that learning does not occur in tasks like the RPM (Alderton & Larson, 1990; Sternberg, 2002), others have presented evidence that individuals show learning effects after both retesting (Bors & Vigneau, 2001) and within a single administration (Bui & Birney, 2014; Verguts & De Boeck, 2002). Verguts and De Boeck found that when solving RPM items, participants tended to use rules that they had previously encountered in the test, thus implying at least some learning. Ren, Wang, Altmeyer, and Schweizer (2014), using fixed-link SEM models (Schweizer, 2006), separated out learning processes from performance (i.e., reasoning) processes in the RPM, showing that item-order has a significant association with performance and that this item-to-item learning process accounted for a substantial proportion of the remaining systematic variance in RPM scores after item-difficulty (reasoning) had been considered.

In essence, and challenging common uni-dimensionality assumptions (Birney & Sternberg, 2006), these are two factors that reliably capture individual differences in RPM performance, but do so, we suggest, by acting across different levels of the test. The first, uncontroversial factor is an ability factor (*Gf* in this case) that is stable within an individual but differs between individuals. The second factor is an experiential factor associated with change as one progresses through the test. This second factor may still be *Gf*, albeit instantiated differently, but it may also be something qualitatively distinct, for instance, attention (Ren et al., 2012) or even impulsivity (Lozano, 2015; Ren, Gong, Chu, & Wang, 2017). In either case, it is a within-person factor operating at, or more accurately, emerging across, items.

Following the goal of Ren et al. (2014), we aim to separate the role of learning from performance within RPM, but do so using a multi-level modeling (MLM) approach (rather than SEM) and with a broad array of cognitive and personality moderators in a sample of high-functioning working adults. First, controlling for item-to-item experience (i.e., item-order), we conceptualize *psychometric complexity* (ψ_C) as a statistical moderation of the cognitive demand of items on performance trajectories (what Ren et al., 2012, refer to as the ability-specific component of fluid reasoning). Second, controlling for item-to-item difficulty, we conceptualize *psychometric learning* (ψ_L) as a statistical moderation of item experience on performance trajectories (Ren et al. refer to this as the position-specific component of fluid reasoning). In the following sections we explicate potential personality and metacognitive moderators of these relationships.

Broader determinants of RPM performance: Cognition-Personality Links and Self-Regulation

Beyond cognitive ability, RPM performance trajectories can be characterized by both task-relevant factors (e.g., emerging knowledge of rules) and task-irrelevant factors (e.g., evolving confidence in one's capacity to perform). Task-relevant learning is germane and intrinsically connected to performance (Sweller, Van Merriënboer, & Paas, 1998), and task-relevant reflection on reasoning is typically considered to facilitate this (Mitchum, Kelley, & Fox, 2016). On the other hand, task-irrelevant learning and task-irrelevant, or self-reflection introduce extraneous factors that may negatively impact performance through influencing metacognitive processes that drive motivation, engagement, effort and sensitivity to changing task demands (e.g. Bouffard, Boisvert, Vezeau, & Larouche, 1995; Heslin, Latham, & Vandewalle, 2005; Pintrich, 2000). Some of these factors are associated with stable personality traits that have been directly linked with cognitive abilities. We consider these cognition-personality links now.

Personality and cognitive ability

In a large sample of 2317, Soubelet and Salthouse (2011) investigated cognition-personality relations as a function of specific cognitive abilities (*Gf*, *Gc*, Memory and Speed) and age (18-96). Focusing here on *Gf*, the highest observed association, as indicated by standardized regression coefficients, was with Openness/Intellect (~ .40). The remaining associations were considerably smaller. The associations with Extraversion, Neuroticism, and less reliably, Agreeableness, were statistically significant at around -.20, -.15, and -.10 (respectively). Conscientiousness < -.10 was not associated with *Gf*. These are remarkably consistent with findings from an earlier meta-analysis by Ackerman and Heggestad (1997).

There have been numerous attempts to provide causal explanations for the cognition-personality links (Ackerman, 1996; Cattell, 1987; Zimprich, Allemand, & Dellenbach, 2009). As they have been found to be reliably associated with cognitive performance, we focus on Openness, Extraversion and Neuroticism, because these three personality factors are most reliably associated with cognitive performance.

Openness/Intellect: Investment of cognitive resources in learning and problem-solving requires facilitating personality traits, dispositions, and interests to support ongoing engagement and motivated self-regulation (Ackerman & Beier, 2003; Beier & Ackerman, 2001; Goff & Ackerman, 1992). Building further from such accounts, Ziegler, Danay, Heene, Asendorpf, and Buhner (2012) introduced the Openness-Fluid-Crystallized-Intelligence (OFCI) process model. The OFCI model attempts to more fully integrate a role of Openness in the broader cognitive developmental pathway. They present evidence in favor of the view that people high in Openness are more likely to be attracted to more learning opportunities which positively affect *Gf*; that *Gf* positively affects Openness because the skills afforded by *Gf* are more likely to lead to success in novel, complex situations; and consistent with Cattell's (1987) *Gf*-*Gc* investment theory, Openness has an indirect effect on *Gc* via a direct effect on *Gf;* though this is not without controversy (see Kan, Kievit, Dolan, & Van Der Maas, 2011).

Extraversion: Deyoung, Peterson, and Higgins (2005) refer to the "tendency to engage actively and flexibly with novelty" as a defining commonalty between Extraversion and Openness/Intellect. They have linked Openness with a conative disposition toward novelty (e.g., intellectual curiosity) and Extraversion as realizing this curiosity in actual exploratory behaviors. Positive associations between these factors are not new. Building from Digman (1997) two-factor models of personality, (e.g., Deyoung, 2006) have proposed two-factor models of personality, where Extraversion and Openness defined one higher-order *plasticity* factor, and Neuroticism (reverse coded), Agreeableness and Conscientiousness defined the second *stability* factor. In spite of the positive correlation between these factors, Extraversion is negatively associated with cognitive performance and Openness positively associated. The patterns of correlations are typically associated with

Neuroticism: Under challenging task demands, high levels of Neuroticism have been associated with negative affective responses including heightened stress, anxiety and self-consciousness (Szymura, 2010). In application, intelligence tests are typically considered to be high-stakes and thus the requirement to complete one has the potential to be viewed by individuals as a performance context (Kozlowski & Bell, 2006). For some individuals, such contexts have been associated with increased anxiety (i.e., test anxiety) and choking under pressure, and both have been shown to have deleterious effects on performance (Smeding, Darnon, & Van Yperen, 2015). Moutafi, Furnham, and Tsaousis (2006) experimentally manipulated the stakes of RPM performance (via whether performance was anonymous or not) and thus test anxiety, and found the Neuroticism-Intelligence relationship to be completely mediated by test anxiety.

Low levels of Neuroticism have also been associated with poor performance but for different reasons. According to Szymura, and echoing Eysenck (1967), poor performance is in part a result of lower arousal. This is because low arousal is thought to limit access to attentional resources, dampen effort, and consequently lead to inadequately monitoring the quality of one's performance. Moderate levels of Neuroticism appear to be most effective for cognitive performance. This non-linear influence of Neuroticism on performance was described by Yerkes–Dodson (1908) who suggested the inverted-U relation is due to within-person differences in the subjective experience of arousal when dealing with cognitive tasks of differing complexity. In a population of individuals with moderate trait levels of Neuroticism, variations in task demands are thought to lead to variations in state arousal, that when effectively regulated, lead to optimal effort and resource allocation, and a positive association with performance (Beckmann, Beckmann, Minbashian, & Birney, 2013; Kahneman, 1973; Szymura, 2010).

Self-regulation of affective and cognitive resources brings us to the role of metacognition and the impacts of being asked to reflect on the correctness of one's responses.

Metacognition and Experience

Using a self-regulatory, metacognitive framework, one can conceptualize cognitive abilities and personality factors driving effort and persistence through a range of cognitive and affective reactions to variations in situational demands (Bandura, 1997; Beckmann, Wood, & Minbashian, 2010; e.g., Kanfer & Ackerman, 1989; Mischel & Shoda, 1995). Planning and monitoring are common features of self-regulation theories regarding the allocation and deployment of cognitive resources to learning and performance. Such metacognitive processes have been shown to underlie an individual's evaluation of their belief in the correctness of their responses, as operationalized by confidence ratings (Stankov & Lee, 2014). Although individual differences research often considers confidence to be a stable trait-like factor that facilitates cognitive performance (Kleitman & Stankov, 2007; Pallier et al., 2002), confidence has also been shown to vary considerably and dynamically from item-to-item (Ackerman, 2014).

Confidence ratings represent a formalized requirement for the participant to reflect on their performance immediately after responding. This monitoring may have a dynamic effect on performance on subsequent items as a test taker gains greater insight into the effectiveness of the reasoning strategies used in previous items, and then uses this information to inform approaches to later items. In this way, self-regulation facilitates task-relevant learning. Metacognition-focused interventions have been shown to improve performance by prompting greater self-assessment and reflection (Azevedo, 2005; Boulware ‐ Gooden, Carreker, Thornhill, & Joshi, 2007; Kramarski & Mevarech, 2003; Kuiper, 2002). However, there is another possible outcome to consider. For some individuals, being required to reflect on the correctness of one's responses might trigger task-irrelevant processing, leading to anxiety and self-doubt, reduced self-confidence, and maladaptive behaviors that can have a negative impact on self-regulation and subsequent performance (e.g.

Bouffard et al., 1995; Heslin et al., 2005; Zimmerman, 2000). This is thought to happen, at least in part, because attentional resources are devoted to thoughts unlikely to facilitate learning, such as ruminating about others' evaluation of their performance (Steele-Johnson, Beauregard, Hoover, & Schmidt, 2000) and the stakes of the assessment (Moutafi et al., 2006). One would expect these to be more marked for those higher in Neuroticism.

Aims and Hypotheses

The first goal of the current study is to investigate the personality correlates of item-to-item performance and learning trajectories in the RPM, particularly in regard to Openness (Ziegler et al., 2012), Extraversion (Deyoung et al., 2005) and Neuroticism (Beckmann et al., 2013; Szymura, 2010). The second, related aim is to consider how metacognitive reflection, operationalized through the requirement to provide confidence ratings, might moderate any observed cognition-personality relations (Ackerman, 2014). The implicit assumption in individual differences research is that requiring confidence ratings does not have a significant impact on performance, however there are reasons why this may not be the case (Mitchum et al., 2016). Our explicit expectation is that given what is known about self-regulatory processes, the requirement to reflect on performance in sufficient detail to provide confidence ratings will facilitate deeper metacognitive processing and ultimately better performance (Boulware‐Gooden et al., 2007; Kuiper, 2002). An alternative account would be that confidence ratings prime some individuals to maladaptively focus on task-irrelevant aspects of performance. This would lead to a negative impact on self-regulation, and ultimately a decline in performance and learning outcomes (e.g. Bouffard et al., 1995; Heslin et al., 2005; Vandewalle, 1997). In the study presented here, we aim to provide an alternative and broader approach to considering the item-position effect investigated by Schweizer and colleagues (e.g., Ren et al., 2014; Schweizer, 2006).

Method

Participants

The participants were 311 mid-level managers from four large international companies based in Australia who were participating in a 5-module leadership training and development program. Of these, 252 participants (age: 34.51 yrs, SD = 6.74yrs; 38% female) had full data and were included in the analyses. The data analyzed were from 77 participants who worked for a major financial institution, 86 for an international airline, 61 for an insurance company, and 28 for a broadcasting company. High school was the highest level of education for 12% of participants, 44% had an undergraduate degree, and the remaining participants held higher degrees (e.g., Masters/PhD). On average, participants had spent 2.06 years (SD = 1.97) in the current job (range = 0.50 – 10.00; median = 1.50 years) and had 5.54 years (SD = 4.61) of management experience overall (range = 0.5 – 21.00; median = 4.00 years).

Design

A mixed-level design was employed. The between-subjects factor was RPM administration format (with or without item-based confidence ratings). The within-person component entails item-to-item performances across the 36 RPM items. Individual differences variables were assessed at baseline approximately 6 months prior to the RPM administration. Participants were assigned with their cohort to complete the RPM either with confidence ratings (6 cohorts, N = 81) or in the standard way without confidence ratings (16 cohorts, N = 171) in as balanced a way as was possible within the context of the project (other aspects of the program remained the same).

Materials

Experimental Task

Raven's Advanced Progressive Matrices - Set II

Each of the 36 RPM items consists of a 3 by 3 matrix of geometric patterns with the bottom right element missing. The task is to identify the rule or rules that govern the horizontal as well as vertical sequence of elements within a matrix moving, and then to identify the element that correctly completes the pattern from 8 options.

Standard Administration: The standard administration requires participants to complete as many items as they can within 40mins. Items not reached were scored as incorrect.

Confidence Administration: A subset of the sample completed confidence ratings after responding to each RPM item. Participants allocated to this condition were asked: “How confident are you that your answer was correct?” Participants responded on a sliding 0-100 scale from “not at all confident” to “extremely confident”. The time to consider and provide confidence ratings for each RPM item was not included in the 40min time limit.

Reasoning Measures

Verbal Reasoning – VR (SHL-VMG4) is a 48-item commercially sourced test that measures the ability to understand and evaluate the logic of various verbal arguments relevant to managerial work (www.shl.com). The task was to decide whether a statement made in connection with the given information was true, untrue, or whether there was insufficient information to judge (Cronbach $\alpha = .82$).

Numerical Reasoning – NR (SHL-NMG4) is a 35-item commercially sourced test that measures the ability to make decisions or inferences based on numerical data and was designed to apply to a range of management level jobs (www.shl.com). The task was to interpret data and combine information from different sources in order to answer the questions given. Calculators were provided so that the emphasis in this test was on understanding and evaluation rather than on computation (Cronbach $\alpha = .91$).

Abstract Reasoning – AR (SHL-DC3.1) is a 40-item commercially sourced test that measures the ability to reason with abstract figures and requires the recognition and application of logical rules governing sequence changes (www.shl.com). The abstract reasoning test consisted of a series of diagrammatic sequences. The task was to identify the underlying structure of this sequence, and select the figure that best completed the pattern (Cronbach α = .85).

SHL Reasoning Ability: Principal components analysis of SHL reasoning measures was used to derive a single measure. The principal component captured 64.71% of the common variance in SHL test scores. Component loadings were *VR* = .768, *NR* = .878, and *AR* = .762.

Personality measures

The Big-Five personality items were assessed using Goldberg's 50-item self-report version (IPIP, see http://ipip.ori.org/ipip), which has 10 items for each factor scored on 0-100 scale anchored by the labels "very inaccurate" to "very accurate". *Openness* (Cronbach α = .78) reflects a continuum of appreciation for intellect and novelty of experience; *Conscientiousness* (Cronbach α = .87) reflects a continuum of orderliness, self-discipline, and control; *Extraversion* (Cronbach α = .88) reflects an engagement with the external world vs independence of the social world; *Agreeableness* (Cronbach α = .76) reflects trusting, generosity vs skeptical, self-interest; and *Neuroticism* (Cronbach α = .85) reflects negative reactivity vs emotional stability.

Procedure

Participants completed the assessments as part of a 5-module leadership training program conducted over 2 years (see Appendix 1). SHL reasoning tests and other measures were administered in Module 1, the RPM was administered in Module 2 about 6 months later. All tests were administered via computer and all sessions were proctored by trained researchers.

Results

Preliminary Checks

As described previously, it was not possible to allocate participating industry managers to conditions completely at random. Differences at baseline in 5 of the 9 variables were observed (see Table A2, Appendix 2). Relative to the Standard group, the Confidence group scored significantly higher on SHL Verbal and Numerical Reasoning, Agreeableness and Openness, and were lower in Neuroticism (all $ps < .05$). There were no significant differences between the groups in terms of SHL Abstract Reasoning (the ability closest in form to RPM), age, gender, employment experience, Extraversion, and Conscientiousness. Differences are considered to be a function of the companies participants came from and the cohorts that their HR managers allocated them to[1]. Given we are investigating performance on the RPM, the prototypical indicator of *Gf*, we explored the extent to which the relationship between SHL reasoning ability (latent factor from the 3 *SHL* tests) and each of the 5 personality measures for each of the 22 cohorts differed by group (with/without confidence ratings). There were no significant differences observed across groups ($.174 < ps < .938$).

Overview of Analyses

The main analyses are reported in two sections. Section 1 describes the derivation of calibrated item-level difficulty parameters which serve as the basis of the trajectory analyses of Section 2. Section 2 provides a simultaneous analysis of the item-by-item performance and learning trajectories as a function of the moderators and experimental condition. Using a multi-level-modeling approach, we consider each moderator separately with its respective group

[1] Although we requested that companies provide a broad range of employees to participate in the program, ultimately the researchers had little control over the final cohort provided.

interaction after controlling for SHL reasoning ability (general cognitive ability). For completeness, an additional set of analogous moderator analyses considers actual confidence ratings provided by the participants.

Section 1: Item Calibration Analyses

The progressive increase of RPM item complexity associated with the administration order introduces a confound when estimating item-difficulty. To deal with this our analytic framework considers both item-difficulty and item-order. This is important statistically (to address the confound) but also important conceptually in terms of moderation effects (psychometric complexity and learning). Item-difficulty was calibrated using the Rasch measurement model as implemented in *Winsteps* (Linacre, 2012); the scale of the metric was specified to UMEAN=0 and USCALE=1. With appropriate fit, item-difficulty parameters are considered to be scale-free estimates independent of the sample on which they are based (Perline, Wright, & Wainer, 1979). Item fit to the Rasch model was satisfactory, as might be expected from an extensively validated assessment (*Model Item parameters*: *RMSE* = .194, *mean* = 0.00, *SD* = 1.87[2], *reliability* = .99, *separation* = 9.59, *infit mean* = 0.98, SD = 0.09, *min* = 0.82, *max* = 1.29). All items were within acceptable fit ranges (*infit mean square* > 0.60 and < 1.40).

Analysis of person fit, however, indicated a relatively high proportion of misfit in the measurement of person ability (11.64% of participants having infit values > 1.40), suggesting item response patterns inconsistent with the Rasch measurement model (*Model Person parameters*: *RMSE* = .476, *mean* = 0.889, SD = 1.12, *reliability* = .82, *separation* = 2.13, *infit mean* = 0.99, SD =

[2] We note that this is relatively large given USCALE (SD) was set to 1. Reanalysis of previously published RPM data (Birney & Bowman, 2009) using the same specifications showed similar fit in a sample of 175 university students (*Item parameters*: *RMSE* = .242, *mean* = 0.00, *SD* = 1.78, *reliability* = .98, *separation* = 7.26, *infit mean* = 0.99, SD = 0. 10, *min* = 0.84, *max* = 1.27), thus we suspect this to be a function of the RPM, not this sample per se.

0.30, *min* = 0.41, *max* = 2.04). An analysis of differential item functioning across condition indicated that only 2 of the 36 items had statistically significant DIF – items 33 and 34 were significantly more difficult when presented with confidence ratings than without, *ps* < .05. Exploring this further in terms of differential test functioning (DTF), a comparison of the observed empirical relationship between item estimates in the two conditions and the identity relationship (which assumes no differences between conditions) revealed that the observed empirical slope is not significantly different from the identity slope, $t_{34} = 0.58$, $p = .282$. As such, we can surmise that because significant DIF is present in only 2 of the 36 items and that DTF is not significant, it is likely that being required to make confidence ratings is probably not responsible for person misfit. While it is also possibly the case that confidence ratings are not having an *overall* differential effect on *mean* test performance, we note that given the confidence group did demonstrate statistically higher SHL reasoning ability, they may have under-performed on the RPM. A linear regression analysis suggests this to be the case. When SHL reasoning ability was controlled for, there was a significant group difference in the Rasch calibrated RPM score, with the Confidence group performing more poorly (standardized $\beta = -0.13$, $t(248) = -2.64$, $p = .009$). This is a difference of approximately 1.35 items.

Section 2: Trajectory Analyses.

The following MLM analyses were conducted using HLM 7 (Raudenbush, Bryk, & Congdon, 2011). An analysis of the variability at each level (unconditional model) indicated that 93.6% of the variability in the data occurs at level 2 (between persons) and 6.4% at level 1 (within person; or equivalently, differences between items). Item-based RPM performance is a binary variable (accuracy = 0/1), and therefore the appropriate underlying model is a logistic one. Item-level variables entered at level 1 were item-order (centered; -17 to 17), item difficulty (Rasch scaled estimates from Section 1; mean = 0), and for those in the Confidence group, confidence ratings (0-100; person-centered). The individual differences variables (grand-mean centered) and experimental group are considered at level 2. Model 1 is represented in Table 1.

The objective of Model 1 is to simultaneously consider the association between performance and item-difficulty (as calibrated in the Rasch analysis of Section 1) controlling for item-order, and the association between performance and item-order controlling for item-difficulty. Inclusion of the moderators and confidence rating group at level 2 provides tests of cross-level interactions with the level 1 parameter estimates – mean RPM performance (intercept), item-difficulty performance trajectories (item-difficulty slopes), and item-order performance trajectories (item-order slopes). Together, this analysis tests whether mean RPM performance and trajectories differ as a function of 1) specific personality factors, 2) which group participants are in (i.e., with or without confidence ratings), and 3) whether there is a higher-order interaction effect between personality and group on performance. Because of the significant differences between groups at the outset in reasoning ability, and because we are interested in the incremental prediction of personality factors beyond cognition, SHL reasoning ability is included as a covariate in all models.

Table 1.

MLM model of the moderator analysis of RPM performance (Model 1).

Level 1:

$$\text{Prob}(Y_{ti} = 1 \mid \pi_i) = \phi_{ti}$$
$$\text{Log}[\phi_{ti} / (1 - \phi_{ti})] = \eta_{ti}$$

$$\eta_{ti} = \pi_{0i} + \pi_{1i}.\text{DIFFICULTY}_{ti} + \pi_{2i}.\text{ORDER}_{ti}$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01}.\text{SHL} + \beta_{02}.\text{group} + \beta_{03}.\text{IV} + \beta_{04}.\text{IV.group} + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11}.\text{group} + \beta_{12}.\text{IV} + \beta_{13}.\text{IV.group} + r_{1i}$$
$$\pi_{2i} = \beta_{20} + \beta_{21}.\text{group} + \beta_{22}.\text{IV} + \beta_{23}.\text{IV.group}$$

Where Y = RPM Accuracy (0/1) for individual i, at item t; DIFFICULTY = Rasch calibrated difficulty; ORDER = centered item order (-17 to 17); Group: 0 = Standard, 1 = Confidence; SHL = SHL reasoning ability; IV = moderating independent variable (SHL, Openness, Extraversion, Neuroticism, Agreeableness, Conscientiousness)

Before reporting the results of these moderator analyses, we wish to draw attention to the fact that a relatively large number of main-effects and cross-level interaction effects are considered. It is thus prudent to reflect on the implication of this in terms of power and alpha inflation for the families of comparisons. The MLM approach is distinctly advantageous in this regard (Gelman, Hill, & Yajima, 2012) compared to OLS regression (Brunner & Austin, 2009). MLM uses a partial pooling process (often referred to as “shrinkage”) that serves to shift parameter estimates and their associated standard errors toward mean coefficients in the complete data. This processes has the desirable effect of shrinking coefficients that are estimated with small accuracy more so than those estimated with higher accuracy (Hox, 2010), thus intervals for comparisons are more likely to include zero (Gelman et al.).

While the MLM approach produces conservative estimates that are more likely to be valid (see Gelman, et al., 2012, and Hox, 2010, for more detailed discussion), we take the following additional steps to strengthen our confidence in the results. First, the effects reported are after controlling for appropriate covariates (a common strategy to increase power). Second, follow-up simple-slopes analyses are only considered when a moderation effect is significant. Finally, we point out for the interested reader that if we had not used an MLM approach, the critical value (for results reported in Table A3) for a standard Bonferroni-adjusted family-wise error rate is t = 2.50, p = .013 (with k=4 family-wise comparisons each for analyses of the intercept, and item-difficulty and item-order slopes).

Moderator Analyses

The first variant of Model 1 tested the reduced base model and included only SHL reasoning ability and Group. It serves as a comparison to subsequent models (i.e., no IV moderator was included). SHL reasoning ability (β = 0.646, t_{245} = 13.65, p < .001) and Group (β = -.261, t_{245} = -2.62,

p = .009) were significant predictors of mean RPM performance (Deviance = 24182.58; Number of parameters = 8). The base model is reported in Table A3.

Six separate analyses based on Model 1 were then run, one for each of the moderators (SHL reasoning ability and the five personality factors). Full details of these analyses are reported in Appendix 3, including 95% confidence intervals on the odds-ratio estimates. Cross-level interaction plots and simple-slopes analyses are provide to support interpretation of the significant moderated associations. Controlling for SHL reasoning ability on mean RPM performance, the addition of Group with each of the personality factors in separate analyses all resulted in significant model improvement (as indicated by significant changes in Deviance estimates relative to the base model, see Table A3). Significant moderation was observed for Openness, Neuroticism, and not for Conscientiousness, as expected. Counter to expectations, significant moderation was observed for Agreeableness, but not for SHL Reasoning ability or Extraversion.

SHL Reasoning Ability: The Model 1 analysis with SHL reasoning ability as the moderator revealed that neither item-difficulty nor item-order trajectories were associated with SHL Reasoning ability. As such, there is no evidence of statistical moderation in terms of a change in the association between performance and cognitive ability as a function of item-difficulty (what Birney and Bowman, 2009, referred to as *psychometric complexity*), and no analogous item-order moderation.

Openness: The Openness factor was a significant predictor of mean RPM score (β = 0.009, t_{243} = 2.45, p = .015), but did not statistically moderate the item-difficulty trajectory. When Openness was controlled for, there were significant group differences in the item-order trajectory (β = -0.039, t_{8428} = -1.98, p = .047). Thus, while there was no direct evidence for a moderating association between Openness and RPM performance trajectories, there was an indirect association. Test takers in the standard administration group had a more positive association with performance as a function of item-order than those who were required to provide confidence ratings (Figure 1).

Insert Figure 1 here

Figure 1. Moderation of item-order trajectory, controlling for Openness, for the RPM standard administration (standard) and RPM administration with confidence ratings (confidence).

Neuroticism: Group membership statistically moderated the association between mean RPM score and Neuroticism (β = 0.015, t_{243} = 2.27, p = .024). After controlling for SHL Reasoning Ability, Neuroticism was more positively, and according to simple-slopes analyses (β = 0.014, t_{75} = 2.48, p = .015), significantly associated with mean RPM performance in the confidence group than in the standard group. Simple-slopes analyses revealed that Neuroticism did not predict mean RPM performance in the standard group (β = 0.001, t_{75} = -0.165, p = .869). Neuroticism also statistically moderated item-difficulty and item-order trajectories (Figure 4A). Higher Neuroticism was associated with less pronounced decline in item-difficulty trajectories (β = 0.007, t_{244} = 2.36, p = .019), and a more pronounced decline in item-order trajectories (β = -0.002, t_{8428} = -2.67, p = .008). This is consistent with an interpretation that the impact of item-difficulty on performance was less marked for those higher in Neuroticism (Figure 2A), and that the impact of item-order was more marked for those lower in Neuroticism (Figure 2B).

Insert Figure 2A here	Insert Figure 2B

Figure 2. Moderation of RPM item-difficulty (A) and item-order (B) trajectories by Neuroticism (± 1.5sd)

Agreeableness: Given Agreeableness has consistently been shown to have little to no association with cognition, no specific predictions were made. However, in the current data, statistically significant associations were observed, although these emerged only as part of the group moderation analysis. Group (standard vs. confidence) statistically moderated the association between mean RPM score and Agreeableness (β = -0.015, t_{243} = -2.06, p = .040). After controlling for SHL reasoning ability, Agreeableness was more negatively, and according to simple-slopes analyses (β = -0.015, t_{75} = -2.455, p = .017), significantly associated with mean RPM performance when confidence ratings were required. There was no significant association between Agreeableness and mean RPM performance under standard conditions (simple-slopes analysis: β = 0.004, t_{167} = 0.933, p = .352). Group membership also moderated the association between Agreeableness and item-difficulty trajectories (β = 0.021, t_{244} = 2.07, p = .039). Higher Agreeableness was associated with a more pronounced positive association with item-difficulty trajectories in the confidence group (Figure 3A), than in the standard group (Figure 3B). The simple-slopes analyses for each group indicated that while the associations were in the opposite directions (as can be observed in the plots), neither was statistically significant in its own right. Stated differently, higher agreeableness was associated with a more pronounced decline in performance as a function of item difficulty under standard conditions, than under confidence conditions. Thus there is evidence for an association between psychometric complexity and Agreeableness but its nature is statistically weak and qualitatively differs depending on whether confidence ratings were required.

Insert Figure 3A here	Insert Figure 3B here

Figure 3. Moderation of item-difficulty trajectory by Agreeableness (± 1.5sd) in Standard (A) and Confidence (B) groups

Conscientiousness and Extraversion: There were no significant unique incremental associations (main-effects or item trajectories) observed for Conscientiousness and Extraversion. Thus, contrary to expectations, Extraversion did not predict mean RPM performance, did not statistically moderate item-difficulty or item-order trajectories, and there were no significant interactions with group. Similarly, when Conscientiousness was included, no statistical moderation was observed and group differences were no longer uniquely significant (see Table A3).

Alignment between confidence and performance

So far we have considered metacognition in terms of between-group differences in being required to provide confidence ratings or not. It is of further interest to investigate whether the confidence ratings themselves moderate the cognitive-personality associations with RPM means and item trajectories. The Model 1 analyses were repeated with the addition of item-level confidence-ratings person-centered at level 1 with the N = 80 participants who provided confidence ratings. When confidence is centered at the level of the individual, the accuracy-confidence slope defines the degree of metacognitive alignment taking into consideration individual differences in use of the 100-point scale. That is, mean confidence is relative to the individual rather than the whole sample. The analyses indicated that overall, the relationship between confidence and performance was positive and significant (β = 0.031364, t_{77} = 11.606, p < 0.0001) suggesting that test takers have overall a realistic appreciation of their performance level. Reasoning ability as measured via SHL was not associated with accuracy-confidence alignment (i.e., not supporting a Dunning-Kruger effect). The only personality factor associated with accuracy-confidence alignment was Extraversion (Figure 4). Higher Extraversion was associated with lower alignment (β = -0.0003, t_{76} = -2.19, p = .032). No other significant associations with RPM performance were observed.

Insert Figure 4 here

Figure 4. Alignment (association between confidence rating and accuracy) interaction with Extraversion

For completeness, a final set of analyses was conducted regressing SHL reasoning ability and each personality factor separately on the actual confidence rating (i.e., with confidence rating as the dependent variable). Item accuracy (β = 18.60, t_{76} = 10.40, p < .001), item difficulty trajectories (β = -3.86, t_{76} = -6.53, p < .001), and item-order trajectories (β = -0.41, t_{76} = -4.26, p < .001) were all significantly associated with item confidence. However, neither SHL reasoning ability nor any of the personality variables significantly predicted mean confidence, nor did they predict item-difficulty and item-order confidence trajectories. Thus whilst there is some evidence to suggest that the requirement to provide confidence ratings has some association with performance, and that this is statistically moderated by personality traits, the actual confidence ratings provided seem to be unrelated with these same traits.

Discussion

In the current work we conceptualize an interplay between cognitive and non-cognitive factors that goes beyond total test scores. Although personality and metacognitive factors have consistently been observed to be associated with intellectual performance (Ackerman & Heggestad, 1997; Soubelet & Salthouse, 2011), there is little research into their associations with item-to-item performance trajectories in cognitive tasks. Our manipulation to require confidence ratings in a sub-sample of the participants served two purposes. First, it allowed us to evaluate whether confidence ratings would have an impact on RPM performance and item-to-item trajectories, and second, by requiring participants to externalize metacognitive processing, we emulated a situation that brought self-regulatory processes to the fore.

Our findings are summarized in two parts. First conceptually, in terms of cognitive-personality associations with performance, and the influence metacognitive reflection has on these associations. And second methodologically, in terms of the implications of considering moderation of individual differences in within-task performance trajectories, which we conceptualize as *psychometric complexity* (ψ_C) and *psychometric learning* (ψ_L).

Cognition-Personality moderation

Based on prior cognition-personality research, we expected three of the five personality factors to be related to overall mean and item-to-item RPM performances: (1) Openness, because it is most commonly found to be associated with cognition (Ackerman & Heggestad, 1997; Soubelet & Salthouse, 2011), and recent research has formalized developmental pathways between Openness, *Gf* and *Gc* (Ziegler et al., 2012). (2) Extraversion, because it has been thought to drive actualization of the Openness disposition (Deyoung et al., 2005), and (3) Neuroticism, because of its role in arousal-based theories of attention (Kahneman, 1973; Szymura, 2010) and test anxiety (Moutafi et al., 2006; Smeding et al., 2015). We did not expect to find consistent cognition-personality associations with Agreeableness and Conscientiousness.

In our study, Openness was related to overall performance as expected from prior research (Ackerman & Heggestad, 1997; Soubelet & Salthouse, 2011), but there was no evidence that Openness statistically moderated item-difficulty or item-order trajectories directly. Neuroticism was found to be important for mean RPM performance. Overall, higher levels of Neuroticism were associated with higher mean RPM performance in the group that was encouraged to metacognitively reflect on their performance via confidence ratings, but not in the standard group. Neuroticism also statistically moderated item-to-item trajectories. These associations did not differ depending on whether confidence ratings were required or not. After controlling for item-order, higher levels of Neuroticism were associated with better performance overall as items became more difficult. Simultaneously controlling for item-difficulty, higher levels of Neuroticism were associated with lower levels of performance as the test progressed (i.e., by item order). These opposing associations are somewhat counter-intuitive. One might reasonably speculate from this data, that Neuroticism supports performance as items become more difficult, but impedes learning as one progresses through the test. This is potentially consistent with the dual competing actions account of Neuroticism – as simultaneously a propensity for arousal, that at moderate levels facilitates performance (Beckmann et al., 2013; Kahneman, 1973; Szymura, 2010), and for anxiety (i.e., worry and test anxiety, e.g., Moutafi et al., 2006). That is, sufficient arousal is necessary to ensure task focus, engagement, and performance. Worry and anxiety on the other hand is associated with task-irrelevant processing that can impede learning, operationalized here as item-order trajectories after controlling for item-difficulty. That Neuroticism may be associated simultaneously with facilitation of performance and impeding learning has important implications for our understanding of intellectual functioning.

The associations between mean RPM performance and item-difficulty trajectories were statistically moderated by Agreeableness. Higher Agreeableness was associated with a more pronounced decline in performance as a function of item difficulty under standard conditions, than under confidence conditions. In fact the nature of the association was diametrically opposed in the two groups (i.e., interaction). We conclude that there is some evidence for Agreeableness to be associated with RPM performance trajectories albeit statistically weak and qualitatively shifting depending on whether confidence ratings were required. However, based on low and inconsistent effect sizes observed in previous literature (Ackerman & Heggestad, 1997; Soubelet & Salthouse, 2011), we made no predictions as to the cognition-Agreeableness association, and therefore do not speculate further on the basis of these statistical effects, other than to note them for future research.

Finally, contrary to expectations, Extraversion did not statistically moderate mean RPM performance. In the supplementary analyses that included actual confidence ratings at level 1, higher Extraversion was significantly (though weakly) associated with poorer alignment between item confidence and accuracy. Again, as we made no specific predictions for such associations, we do not speculate further, other than to note this as an area for future investigations.

Metacognition

Metacognition is a critical component of self-regulation and has been considered as a bridge linking intellectual abilities and personality (Stankov, 1999). Using a self-regulatory framework, one can conceptualize cognitive abilities and personality factors driving effort and persistence through a range of cognitive and affective reactions to variations in situational demands (Bandura, 1997; Beckmann et al., 2010; Mischel & Shoda, 1995). The increase in item complexity can be seen as a situational change. In the current research we went beyond the analysis of the cognition-personality link that is based on the total performance score. We were interested in whether and how the processes triggered by this situational change in demand is reflected in how item-to-item trajectories of experience are associated with performance outcomes. Of particular interest was whether requiring participants to externalize their metacognitive processing by asking them to reflect and report on their confidence in the correctness of their responses to each of the 36 RPM items, would accentuate or attenuate the cognition-personality associations. Two accounts were proposed. The first suggests that reflecting on one's processing facilitates performance (Boulware ‐ Gooden et al., 2007; Kuiper, 2002). The second suggests that for some test takers, the same reflection leads to maladaptive behaviors and consequentially poorer performance (e.g. Bouffard et al., 1995; Heslin et al., 2005; Vandewalle, 1997).

Group membership (metacognition experimentally encouraged vs. not encouraged) was shown to have a statistically significant effect across several of the associations observed. First, there were group differences in overall RPM performance. Controlling for SHL reasoning ability (i.e., latent general cognitive ability), those required to provide confidence ratings had significantly lower RPM scores. From this perspective, it would seem that (1) performance is not immune to confidence ratings, and (2) that the facilitating effect of self-reflection reported in other work cannot claim generality. Strictly speaking, the methodology employed in this study (i.e., a mixture of experimental and quasi-experimental design) does not support strong causal claims. We will however note possible avenues of future research more suitable to identifying mechanism that underpin the cognition-personality link and its impact on performance.

Neuroticism, because of its associations at all levels of analyses in our data, is the strongest candidate for further investigations. We earlier alluded to both facilitating arousal and impeding anxiety explanations as a plausible mechanism for the observed associations. Such conceptual suppositions do have some empirical support (e.g., Moutafi et al., 2006), but further experimental research is needed. We are of the view that investigations into distinctions between state and trait Neuroticism (Beckmann et al., 2013), possibly using confidence ratings as an experimental catalyst, may prove particularly fruitful in this regard.

Methodological Issues

The multi-level models investigated three levels of effects simultaneously – mean RPM performance, the incremental statistical moderators of item-difficulty trajectories, and the incremental statistical moderators of item-order trajectories. These models are comparable to the fixed-links model discussed by Schweizer and his colleagues (Ren et al., 2012; Schweizer, 2006; Schweizer, Altmeyer, Ren, & Schreiner, 2015; Wang, Ren, Altmeyer, & Schweizer, 2013), and both are consistent with the general class of SEM latent variable models of multilevel data (Muthen, 1997). The main overarching difference is that we model item responses using a multi-level logistic regression, whereas Schweizer models item-clusters using a more standard SEM implementation. In our MLM model, the mean item-difficulty trajectory (see Table 1, π_{1i}) is analogous to the latent ability factor in Schweizer (2006). Instead of item clusters being weighted/constrained to be equal, we weight/constrain each item by its empirical difficulty derived from the IRT calibration analyses[3].

The item-order trajectory (Table 1, π_{2i}) is comparable to Schweizer's item-position latent factor, where each item is weighted by its linear position in the test. Schweizer (2006) has also used other link functions (e.g., quadratic functions). Finally, the intercept in the fixed link model is comparable to the overall intercept in the MLM model (Table 1, π_{0i}), what is in effect the mean RPM score parameter. The final difference is in how we attempt to understand the latent factors. Schweizer and his colleagues (e.g., Ren et al., 2012; Schweizer, 2006; Schweizer et al., 2015; Wang et al., 2013) have explored fixed-links models to partition variance in numerous cognitive functions (e.g., attention, WM, and learning), whereas we have simultaneously explored general reasoning ability (as a covariate) and moderating factors in the personality realm, while also introducing an experimental catalyst to prime self-reflection, which in practice seems to have accentuated associations. It is interesting to note, that in our data, item-order effects did not emerge alone as in Schweizer's work, but were only present in association with personality moderators. Thus we were unable to fully replicate Schweizer's item-position effects in our cohort of senior managers.

It is also important to be cognizant of the fact that the amount of systematic within-subject, item-to-item variability relative to between-subject variability in performance on RPM items is small – 6.4% in total. This is consistent with what other researchers have reported (Ren et al.,

---

[3] Although defined empirically here, item difficulty in RPM is concordant with theoretical (albeit posthoc) analyses of item induction/WM demand (Carpenter et al., 1990).

2014) and unsurprising given the extensive between-subject validation data that exists for the RPM. Item-difficulty is closely associated with item-order in the RPM (*r* = .95), yet the additional information provided by item-difficulty over and above item-order provides a statistically significant systematic contribution to the explained variability in item performance.

Psychometric Complexity and Psychometric Learning

The progressive increase of difficulty associated with a standard RPM administration order introduces a confound when estimating item-difficulty. To resolve the item order-difficulty confound (at least in part), our analytic framework considers both item-difficulty and item-order simultaneously (as does Schweizer, 2006). This is also theoretically important in terms of the moderation effects that define our broadened conceptualization of *psychometric complexity* and our new conceptualization of *psychometric learning*, to which we now turn.

Psychometric Complexity (ψ_C)*:* Task manipulations that result in monotonic increases in the association between task performance and *Gf* test scores, define *psychometric complexity* (Birney & Bowman, 2009). Psychometric complexity as a specific type of statistical moderation, can also be conceptualized for other attributes that impact on reasoning differentially and monotonically as a function of increased task demand. For instance, Neuroticism has been implicated in cognitive performance (Beckmann et al., 2013; Debusscher, Hofmans, & De Fruyt, 2014). If changes in the relationship between Neuroticism and reasoning performance differed systematically and monotonically across the 36 RPM items as a function of item-difficulty (controlling for item-order), then there would be evidence for a ψ_C effect for Neuroticism. This is what we observed.

Psychometric Learning (ψ_L)*:* We introduce *psychometric learning* as a new, experientially-driven concept analogous to psychometric complexity. The contribution of the ψ_L concept is based on an assumption that variables that statistically moderate individual differences in item-order progression, after controlling for item complexity, are moderators of *learning*. We argue that under these circumstances, the sequential item-to-item progression is an evolving task-specific experiential factor because of the accumulation of experiences associated with progression through the RPM test. Thus, if changes in the relationship between Neuroticism and reasoning performance differed systematically and monotonically across the 36 RPM items as a function of item-order (controlling for item-difficulty), then there would be evidence for a ψ_L effect for Neuroticism. This is what we observed. Higher Neuroticism was associated with a decline in 'learning'. Extending the application of psychometric complexity to psychometric learning, by proposing a theoretical account of the statistical moderation of the trajectory of item-order performance (controlling for item difficulty), moves the discussion away from atheoretical statistical moderation effects, and makes clear the necessity to seek out and experimentally investigate causal pathways. At their core, ψ_C and ψ_L embrace the tradition of experimental research of individual differences, as they are based on an examination of experimental manipulations (item-demand and item-order) and a test of concomitant changes in systematic variance.

Limitations and Future Directions

While we have argued that the findings of this research are compelling, there are several important aspects that limit generalisability and potentially replicability. These include the fact that the population of managers our sample has been drawn from is difficult to circumscribe; implications of the unique personal experiences of the participants is hard to quantify (e.g., some are highly educated, some are not; some were born and raised in Australia, some were not); some small a priori differences between groups were apparent in personality and cognitive constructs; and the professional learning and development context in which these assessments were taken cannot easily be replicated. Yet, these are real people, working within contexts where psychometric assessments are heavily used and highly valued.

The cognition-personality links have often been presented more as static descriptions of observed associations (Soubelet & Salthouse, 2011) rather than as dynamic developmental, causal models. The various explanatory models that have been proposed, of which the work of Ackerman and his colleagues is notable (e.g., Ackerman, 1996; Ackerman & Beier, 2003; Cattell, 1987; Kanfer & Ackerman, 1989; Zimprich et al., 2009), have been based on correlational data. Our data is certainly still predominantly correlational, although we have introduced an experimental manipulation in terms of the requirement to provide confidence ratings. The requirement for some participants to externalize metacognitive processing by reporting confidence in the correctness of their response is not part of the standardized RPM administration. As such, it presents somewhat of an artificial requirement in a "real-world context". However, as a method to further investigate cognition-personality associations, it shows promise. This is particularly so as it may provide a basis to better understand the bridge between personality and intelligence, that Stankov (1999) referred to as the "no man's land". Current work in our lab is investigating manipulations of item-order; working with different populations, and with different metacognitive cues and non-cognitive factors.

Conclusion

Our research extends the cognition-personality associations observed in the individual difference literature to item performance trajectories. Moving to the level of item responses and introducing experimental manipulations (e.g., confidence ratings) that serve as catalysts for modification of cognition-personality relations, provide an important avenue for integrating experimental and differential methods. Performance is not immune to being asked to provide confidence ratings, but in addition to their theoretical importance, they show promise as an experimental tool. Together they may inform causal models of cognitive performance and provide the basis of a more nuanced account of self-regulation of resources important to reasoning and learning. We finish by noting that our work supports the view that the broader context of a cognitive assessment does not remain constant. It is fluid and differentially impacts performance that goes beyond the intellect being measured. We proposed a conceptualization of psychometric complexity and psychometric learning as theoretically derived accounts of statistical moderation that may provide a bridge to causal accounts.

Acknowledgements

This research was supported under Australian Research Council's Linkage Projects funding scheme (project LP0669552) and Discovery Projects funding scheme (project DP140101147). The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. All required ethics approval was obtained through the University of Sydney and University of New South Wales human ethics review committees.

References

Ackerman, P. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Ackerman, P. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227-257.

Ackerman, P., & Beier, M. E. (2003). Trait complexes, cognitive investment, and domain knowledge. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise.* (pp. 1-30). Cambridge, UK.: Cambridge University Press.

Ackerman, P., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219-245.

Ackerman, R. A. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*, 1349-1368.

Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's advanced progressive matrices items. *Educational and Psychological Measurement, 50*, 887-900.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*, 199-209.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational behavior and human decision processes, 50*, 248-287.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Beckmann, J. F., & Birney, D. P. (2012). What happens before and after it happens: Insights regarding antecedences and consequences of adult learning. *Learning and Individual Differences*, 561-562.

Beckmann, N., Beckmann, J. F., Birney, D. P., & Wood, R. E. (2015). A problem shared is learning doubled: Deliberative processing in dyads improves learning in complex dynamic decision-making tasks. *Computers in Human Behavior, 48*, 654-662.

Beckmann, N., Beckmann, J. F., Minbashian, A., & Birney, D. P. (2013). In the heat of the moment: On the effect of state neuroticism on task performance. *Personality & Individual Differences, 54*, 447-452.

Beckmann, N., Wood, R. E., & Minbashian, A. (2010). It depends how you look at it: On the relationship between neuroticism and conscientiousness at the within- and the between-person levels of analysis. *Journal of Research in Personality, 44*, 593-601.

Beier, M. E., & Ackerman, P. L. (2001). Current-events knowledge in adults: An investigation of age, intelligence, and nonability determinants. *Psychology and Aging, 16*, 615-628.

Birney, D. P., Beckmann, J. F., & Wood, R. E. (2012). Precursors to the development of flexible expertise: Metacognitive self-evaluations as antecedences and consequences in adult learning. *Learning and Individual Differences, 22*, 563-574.

Birney, D. P., & Bowman, D. B. (2009). An experimental-differential investigation of cognitive complexity. *Psychology Science Quarterly, 51*, 449-469.

Birney, D. P., Bowman, D. B., Beckmann, J., & Seah, Y. (2012). Assessment of processing capacity: Latin-square task performance in a population of managers. *European Journal of Psychological Assessment, 28*, 216-226.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the Influence of Relational Complexity on Reasoning: The Development of the Latin Square Task. *Educational and Psychological Measurement, 66*, 146-171.

Birney, D. P., & Sternberg, R. J. (2006). Intelligence and cognitive abilities as competencies in development. In E. Bialystok & G. Craik (Eds.), *Lifespan Cognition: Mechanisms of Change* (pp. 315-330). New York: Oxford University Press.

Bors, D. A., & Vigneau, F. (2001). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences, 13*, 291-312.

Bouffard, T., Boisvert, J., Vezeau, C., & Larouche, C. (1995). The impact of goal orientation on self‐regulation and performance among college students. *British Journal of Educational Psychology, 65*, 317-329.

Boulware‐Gooden, R., Carreker, S., Thornhill, A., & Joshi, R. (2007). Instruction of metacognitive strategies enhances reading comprehension and vocabulary achievement of third‐grade students. *The Reading Teacher, 61*, 70-77.

Brunner, J., & Austin, P. C. (2009). Inflation of Type I error rates in multiple regression when independent variables are measured with error. *The Canadian Journal of Statistics, 37*, 33-46.

Bui, M., & Birney, D. P. (2014). Learning and individual differences in Gf processes and Raven's. *Learning and Individual Differences, 32*, 104-113.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404-431.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam: Elsevier Science Publishers.

Crawford, J. D. (1991). Intelligence, task complexity and the distinction between automatic and effortful mental processing. In H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 119-144). Hillsdale, NJ: Lawrence Erlbaum.

Debusscher, J., Hofmans, J., & De Fruyt, F. (2014). The Curvilinear relationship between state neuroticism and momentary task performance. *Plos One, 9*, e106989.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology, 91*, 1138-1151.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of Openness/Intellect: Cognitive and neuropsychological Correlates of the fifth factor of personality. *Journal of Personality, 73*, 825-858.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246-1256.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309-331.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield, NJ: Thomas.
Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 189-211.
Goff, M., & Ackerman, P. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84*, 537-552.
Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 803-831.
Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*, 842-856.
Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). New York: Routledge.
Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.
Kan, K., Kievit, R. A., Dolan, C. V., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence, 39*.
Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 66-71.
Kanfer, R., & Ackerman, P. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657-690.
Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161-173.
Kozlowski, S. W. J., & Bell, B. S. (2006). Disentangling achievement orientation and goal setting: Effects on self-regulatory processes. *Journal of Applied Psychology, 91*, 900-916.
Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American educational research journal, 40*, 281-310.
Kuiper, R. (2002). Enhancing metacognition through the reflective use of self-regulated learning strategies. *The Journal of Continuing Education in Nursing, 33*, 78-87.
Linacre, J. (2012). WINSTEPS Rasch measurement computer program (v3.75.0). Chicago: Winsteps.com.
Lozano, J. H. (2015). Are impulsivity and intelligence truly related constructs? Evidence based on the fixed-links model. *Personality & Individual Differences, 85*, 192-198.
Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107-127.
Minbashian, A., Wood, R. E., & Beckmann, N. (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology, 9*, 793-806.
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268.
Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*, 200-219.
Moutafi, J., Furnham, A., & Tsaousis, I. (2006). Is the relationship between intelligence and trait Neuroticism mediated by test anxiety? *Personality and Individual Differences, 40*, 587-597.
Muthen, B. O. (1997). Laten variable modeling of longitudinal and multilevel data. *Sociological Methodology, 27*, 453-480.
Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgements. *The Journal of General Psychology, 129*, 257-299.
Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237-255.
Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning.
Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*, 41-70.
Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). HLM7: Linear and Nonlinear Modelling. Lincolnwood, IL: Scientific Software International.
Raven, J. C. (1941). Standardisation of progressive matrices, 1938. *British Journal of Medical Psychology, XIX*, 137-150.
Ren, X., Goldhammer, F., Moosbrugger, H., & Schweizer, K. (2012). How does attention relate to the ability-specific and position-specific components of reasoning measured by APM? *Learning and Individual Differences, 22*, 1-7.
Ren, X., Gong, Q., Chu, P., & Wang, T. (2017). Impulsivity is not related to the ability and position components of intelligence: A comment on Lozano (2015). *Personality & Individual Differences, 104*, 533-537.

Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences, 31*, 30-35.

Schweizer, K. (2006). The fixed-links model for investigating the effects of general and specific processes on intelligence. *Methodology, 2*, 149-160.

Schweizer, K., Altmeyer, M., Ren, X., & Schreiner, M. (2015). Models for the detection of deviations from the expected processing strategy in completing the Items of cognitive measures. *Multivariate Behavioral Research, 50*, 544-554.

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin, 137*, 421-442.

Smeding, A., Darnon, C., & Van Yperen, N. W. (2015). Why do high working memory individuals choke? An examination of choking under pressure effects in math from a self-improvement perspective. *Learning and Individual Differences, 37*, 176-182.

Soubelet, A., & Salthouse, T. A. (2011). Personality-cognition relations across audulthood. *Developmental Psychology, 47*, 303-310.

Spearman, C. (1927). The abilities of man. New York: Macmillan.

Stankov, L. (1999). Mining on the "No Man's Land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 315-337). Washington, DC: American Psychological Association.

Stankov, L. (2000). Complexity, metacognition and fluid intelligence. *Intelligence, 28*, 121-143.

Stankov, L., & Crawford, J. D. (1993). Ingredients of complexity in fluid intelligence. *Learning and Individual Differences, 5*, 73-111.

Stankov, L., & Lee, J. (2014). Quest for the best non-cognitive perdictor of academic achievement. *Educational Psychology, 34*, 1-8.

Steele-Johnson, D., Beauregard, R. S., Hoover, P. B., & Schmidt, A. M. (2000). Goal orientation and task demand effects on motivation, affect, and performance. *Journal of Applied Psychology, 85*, 724-738.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J. (2002). Intelligence is not just inside the head: The theory of successful intelligence. *Improving academic achievement: Impact of psychological factors on education*, 227-244.

Sternberg, R. J., & Gastel, J. (1989). Coping with novelty in human intelligence: An empirical investigation. *Intelligence, 13*, 187-197.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Szymura, B. (2010). Individual differences in resource allocation model. In A. Gruszka, G. Matthews & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control* (pp. 231-246). New York, NY: Springer.

VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*, 995-1015.

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices test. *European Journal of Cognitive Psychology, 14*, 521-547.

Wang, T., Ren, X., Altmeyer, M., & Schweizer, K. (2013). An account of the relationship between fluid intelligence and complex learning in considering storage capacity and executive attention. *Intelligence, 41*, 537-545.

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Buhner, J. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality, 46*, 173-183.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology, 25*, 82-91.

Zimprich, D., Allemand, M., & Dellenbach, M. (2009). Openness to experience, fluid intelligence, and crystallized intelligence in middle-aged and old adults. *Journal of Research in Personality, 43*, 444-454.

Appendix 1

At the time of submission, the Accelerated Learning Laboratory had provided leadership training program to students and industry mangers. An analysis of a subsample of the industry manager data is reported in the current paper. At the time of submission, a number of further analyses based on subsets of this pool of participants have also been published, including (Beckmann & Birney, 2012; Beckmann, Beckmann, Birney, & Wood, 2015; Beckmann et al., 2013; Beckmann et al., 2010; Birney, Beckmann, & Wood, 2012; Birney, Bowman, Beckmann, & Seah, 2012; Minbashian, Wood, & Beckmann, 2010). We acknowledge dependencies in reference to any evidence presented in support of theoretical interpretation based on these common participants where they occur.

Appendix 2

Table A2
Descriptive statistics (Mean and SD in parentheses) and Correlations between raw scores for RPM with standard (lower triangle) and confidence rating (upper triangle) administration conditions

		Descriptive Statistics					Correlations								
Moderators		**Standard**		**Confidence**		**Cohen's d**	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RPM Sum Score	(1)	22.54	(5.46)	22.67	(4.64)	.025		**.31**	**.50**	**.50**	-.10	-.17	.15	**-.26**	**-.30**
Verbal Reasoning	(2)	50.24	(9.83)	53.68	(9.15)	**.358**	**.52**		**.51**	**.36**	-.01	-.02	-.03	**-.23**	-.13
Numerical Reasoning	(3)	48.96	(9.49)	54.44	(9.99)	**.568**	**.64**	**.54**		**.54**	-.04	-.04	-.14	-.09	-.03
Abstract Reasoning	(4)	54.74	(10.48)	55.53	(10.72)	.075	**.56**	**.30**	**.54**		-.07	-.04	-.02	-.12	-.13
Openness	(5)	66.40	(13.28)	71.17	(13.02)	**.361**	**.17**	**.24**	-.04	.01		**.22**	-.15	.17	-.06
Extraversion	(6)	60.80	(14.91)	61.23	(17.37)	.027	.04	.07	-.04	-.01	**.33**		**-.31**	.05	**.20**
Neuroticsim	(7)	32.64	(14.59)	28.54	(13.90)	**-.285**	-.10	.02	**-.14**	-.12	-.10	**-.39**		**-.47**	-.18
Agreeableness	(8)	71.61	(10.81)	74.79	(10.07)	**.301**	.00	-.12	-.09	.04	.12	.07	**-.35**		**.24**
Conscientiousness	(9)	69.01	(13.75)	72.59	(14.09)	.258	-.08	**-.14**	-.03	.03	.03	**.17**	**-.36**	**.33**	

Notes: Cohen's d between condition effect size (bold = 95% CI does not contain 0); RPM Standard: $N = 170$, for $r > .125$ $p < .05$; RPM confidence: $N = 81$, for $r > .184$ $p < .05$; bold = $p < .05$.

Appendix 3

Table A3

Moderation analyses of mean, item-difficulty, and item-order trajectories by condition

Fixed Effect: Base Model	β	SE	t	df	p	OR	
Mean RPM							
Mean	**0.860**	**0.053**	**16.16**	**245**	**0.000**	**2.362**	**(2**
Group	**-0.261**	**0.100**	**-2.62**	**245**	**0.009**	**0.770**	**(0**
SHL Reasoning	**0.646**	**0.047**	**13.65**	**245**	**0.000**	**1.907**	**(1**
Difficulty Trajectory							
Mean	**-0.858**	**0.049**	**-17.55**	**247**	**0.000**	**0.424**	**(0**
Order Trajectory							
Mean	0.000	0.008	0.02	8431	0.983	1.000	(0

Base Model: Deviance = 24182.58; Number of parameters = 8

Fixed Effect: SHL Reasoning	β	SE	t	df	p	OR	
Mean RPM							
Mean	**0.802**	**0.044**	**18.24**	**244**	**0.000**	**2.230**	**(2**
Group	-0.118	0.098	-1.20	244	0.231	0.889	(0
SHL Reasoning	**0.710**	**0.052**	**13.66**	**244**	**0.000**	**2.034**	**(1**
Group x SHL Reasoning	0.187	0.106	1.76	244	0.080	1.205	(0
Difficulty Trajectory							
Mean	**-0.871**	**0.053**	**-16.48**	**244**	**0.000**	**0.419**	**(0**
Group	-0.025	0.116	-0.21	244	0.833	0.976	(0
SHL Reasoning	-0.020	0.060	-0.34	244	0.733	0.980	(0
Group x SHL Reasoning	-0.183	0.119	-1.55	244	0.124	0.832	(0
Order Trajectory							
Mean	0.001	0.009	0.13	8428	0.899	1.001	(0
Group	-0.024	0.020	-1.23	8428	0.217	0.976	(0
SHL Reasoning	0.005	0.010	0.44	8428	0.663	1.005	(0
Group x SHL Reasoning	0.035	0.018	1.92	8428	0.054	1.036	(0

Deviance = 24161.14; Number of parameters = 15, $\chi^2(7) = 13.89$, p = .0184

Fixed Effect: Openness	β	SE	t	df	p	OR	
Mean RPM							
Mean	**0.854**	**0.051**	**16.84**	**243**	**0.000**	**2.348**	**(2**
Group	-0.125	0.117	-1.06	243	0.289	0.883	(0
Openness	**0.009**	**0.004**	**2.45**	**243**	**0.015**	**1.009**	**(1**
SHL Reasoning	**0.634**	**0.048**	**13.08**	**243**	**0.000**	**1.885**	**(1**
Group x Openness	-0.017	0.010	-1.72	243	0.087	0.983	(0
Difficulty Trajectory							
Mean	**-0.869**	**0.057**	**-15.13**	**244**	**0.000**	**0.419**	**(0**
Group	0.041	0.118	0.35	244	0.725	1.042	(0
Openness	-0.006	0.004	-1.47	244	0.143	0.994	(0
Group x Openness	-0.005	0.009	-0.55	244	0.581	0.995	(0
Order Trajectory							
Mean	0.008	0.010	0.82	8428	0.412	1.008	(0
Group	**-0.039**	**0.020**	**-1.98**	**8428**	**0.047**	**0.962**	**(0**
Openness	0.000	0.001	0.52	8428	0.604	1.000	(0
Group x Openness	0.002	0.002	1.50	8428	0.133	1.002	(0

Deviance = 24156.37; Number of parameters = 16; $\chi^2(8) = 26.22$, p = .0004

Table A3 (continued)

Fixed Effect: Neuroticism	**β**	**SE**	**t**	**df**	**p**	**OR**	**95% CI**
Mean RPM							
Mean	**0.833**	**0.053**	**15.71**	**243**	**0.000**	**2.299**	**(2.072,2.553)**
Group	-0.096	0.096	-1.00	243	0.318	0.908	(0.751,1.098)
Neuroticism	-0.001	0.003	-0.29	243	0.775	0.999	(0.992,1.006)
SHL Reasoning	**0.648**	**0.047**	**13.82**	**243**	**0.000**	**1.911**	**(1.743,2.096)**
Group x Neuroticism	**0.015**	**0.007**	**2.27**	**243**	**0.024**	**1.015**	**(1.002,1.028)**
Difficulty Trajectory							
Mean	**-0.862**	**0.057**	**-15.04**	**244**	**0.000**	**0.422**	**(0.377,0.473)**
Group	-0.027	0.111	-0.25	244	0.805	0.973	(0.782,1.211)
Neuroticism	**0.007**	**0.003**	**2.36**	**244**	**0.019**	**1.007**	**(1.001,1.014)**
Group x Neuroticism	-0.009	0.007	-1.22	244	0.225	0.991	(0.977,1.005)
Order Trajectory							
Mean	0.009	0.010	0.94	8428	0.345	1.009	(0.990,1.028)
Group	-0.030	0.019	-1.59	8428	0.113	0.971	(0.935,1.007)
Neuroticism	**-0.002**	**0.001**	**-2.67**	**8428**	**0.008**	**0.998**	**(0.997,1.000)**
Group x Neuroticism	0.001	0.001	0.54	8428	0.586	1.001	(0.998,1.003)

Deviance = 24157.39; Number of parameters = 16; $\chi^2(8) = 25.19$, p = .0006

Fixed Effect: Agreeableness	**β**	**SE**	**t**	**df**	**p**	**OR**	**95% CI**
Mean RPM							
Mean	0.834	0.052	16.16	243	0.000	2.302	(2.080,2.549)
Group	-0.110	0.100	-1.11	243	0.270	0.896	(0.736,1.090)
Agreeableness	0.004	0.004	0.84	243	0.405	1.004	(0.995,1.013)
SHL Reasoning	**0.643**	**0.048**	**13.34**	**243**	**0.000**	**1.903**	**(1.731,2.093)**
Group x Agreeableness	**-0.015**	**0.007**	**-2.06**	**243**	**0.040**	**0.986**	**(0.972,0.999)**
Difficulty Trajectory							
Mean	**-0.861**	**0.059**	**-14.57**	**244**	**0.000**	**0.423**	**(0.376,0.475)**
Group	-0.066	0.122	-0.54	244	0.592	0.936	(0.736,1.192)
Agreeableness	-0.009	0.005	-1.92	244	0.055	0.991	(0.981,1.000)
Group x Agreeableness	**0.021**	**0.010**	**2.07**	**244**	**0.039**	**1.021**	**(1.001,1.041)**
Order Trajectory							
Mean	0.009	0.010	0.86	8428	0.389	1.009	(0.989,1.028)
Group	-0.022	0.020	-1.08	8428	0.279	0.978	(0.940,1.018)
Agreeableness	0.001	0.001	1.62	8428	0.105	1.001	(1.000,1.003)
Group x Agreeableness	-0.002	0.002	-1.38	8428	0.169	0.998	(0.995,1.001)

Deviance = 24162.89; Number of parameters = 16; $\chi^2(8) = 19.69$, p = .0042

Table A3 (continued)

Fixed Effect: Extraversion	β	SE	t	df	p	OR	95% CI
Mean RPM							
Mean	**0.831**	**0.052**	**15.98**	**243**	**0.000**	**2.297**	**(2.073,2.544)**
Group	-0.141	0.097	-1.45	243	0.148	0.868	(0.717,1.052)
Extraversion	0.003	0.004	0.73	243	0.469	1.003	(0.995,1.010)
SHL Reasoning	**0.647**	**0.047**	**13.78**	**243**	**0.000**	**1.909**	**(1.741,2.094)**
Group x Extraversion	-0.012	0.006	-1.96	243	0.051	0.988	(0.977,1.000)
Difficulty Trajectory							
Mean	**-0.848**	**0.058**	**-14.51**	**244**	**0.000**	**0.428**	**(0.382,0.481)**
Group	-0.053	0.114	-0.47	244	0.640	0.948	(0.758,1.186)
Extraversion	0.002	0.004	0.64	244	0.525	1.002	(0.995,1.010)
Group x Extraversion	0.003	0.009	0.39	244	0.699	1.003	(0.987,1.020)
Order Trajectory							
Mean	0.006	0.010	0.66	8428	0.511	1.006	(0.987,1.026)
Group	-0.022	0.019	-1.12	8428	0.264	0.979	(0.942,1.016)
Extraversion	-0.001	0.001	-0.88	8428	0.380	0.999	(0.998,1.001)
Group x Extraversion	0.000	0.002	0.11	8428	0.910	1.000	(0.997,1.003)

Deviance = 24163.76; Number of parameters = 16; $\chi^2(8) = 18.82$, p = .0057

Fixed Effect: Conscientiousness	β	SE	t	df	p	OR	95% CI
Mean RPM							
Mean	**0.826**	**0.053**	**15.62**	**243**	**0.000**	**2.284**	**(2.058,2.535)**
Group	-0.088	0.101	-0.87	243	0.385	0.915	(0.750,1.118)
Conscientiousness	-0.004	0.003	-1.17	243	0.245	0.996	(0.990,1.003)
SHL Reasoning	**0.639**	**0.047**	**13.53**	**243**	**0.000**	**1.894**	**(1.726,2.079)**
Group x Conscientiousness	-0.011	0.007	-1.61	243	0.108	0.989	(0.976,1.002)
Difficulty Trajectory							
Mean	**-0.852**	**0.059**	**-14.57**	**244**	**0.000**	**0.426**	**(0.380,0.478)**
Group	-0.076	0.117	-0.65	244	0.515	0.927	(0.737,1.166)
Conscientiousness	-0.006	0.004	-1.48	244	0.140	0.994	(0.986,1.002)
Group x Conscientiousness	0.015	0.008	1.92	244	0.056	1.016	(1.000,1.032)
Order Trajectory							
Mean	0.007	0.010	0.77	8428	0.443	1.008	(0.988,1.027)
Group	-0.020	0.020	-1.01	8428	0.312	0.980	(0.943,1.019)
Conscientiousness	0.001	0.001	1.75	8428	0.081	1.001	(1.000,1.003)
Group x Conscientiousness	-0.002	0.001	-1.85	8428	0.064	0.998	(0.995,1.000)

Deviance = 24159.17; Number of parameters = 16; $\chi^2(8) = 23.42$, p = .0011

Notes: β = unstandardized regression coefficient; OR = Odds ratio; 95% CI = 95% confidence interval on the odds ratio; Group: 0 = RPM standard, 1 = RPM confidence; All moderators are grand-mean centered. Item-Difficulty and Item-Order are centered at 0. Test of deviance is based on a comparison with the reduced base model of Item-difficulty and Item-order at level 1, and condition and SHL reasoning at level. Confidence intervals including 1.0 are suggestive of unreliable associations.